UNP JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 1 Nomor 4 Tahun 2023 Halaman 361 - 368.
ISSN(*Online*): 2985-475X

Comparison of Error Rate Prediction Methods in Binary Logistic Regression Modeling for Imbalanced Data

Bahri Annur Sinaga, Dodi Vionanda*, Dony Permana, dan Admi Salma
Departemen Statistika, Universitas Negeri Padang, Padang, Indonesia
*Corresponding author: dodi_vionanda@fmipa.unp.ac.id

Submitted : 18 Juli 2023
Revised : 11 Agustus 2023
Accepted : 18 Agustus 2023

ABSTRACT

Binary logistic regression is a regression analysis used in classification modeling. The performance of binary logistic regression can be seen from the accuracy of the model formed. Accuracy can be measured by predicting the error rate. One method of predicting the error rate that is often used is cross-validation. There are three algorithms in cross-validation: leave one out, hold out, and k-fold. Leave one out is a method that divides data based on the number of observations so that each observation has the opportunity to become testing data but requires a long time in the analysis process when the number of observations is large. Hold out is the simplest algorithm that only divides the data into two parts randomly, so there is a possibility that important data does not become training data. K-fold is an algorithm that divides data into several groups, but k-fold is not suitable for data that has a small number of observations. In reality, real data is often imbalanced. In logistic regression,when the data is increasingly imbalanced, the prediction results will approach the number of minority classes. This research focuses on the comparison of error rate prediction methods in binary logistic regression modeling with imbalanced data. This study uses three types of data, namely univariate, bivariate, and multivariate, which are generated by differences in population mean and correlation between independent variables.The results obtained show that the k-fold algorithm is the most suitable error rate prediction algorithm applied to binary logistic regression.

Keywords:*Binary Logistc Regression, Hold Out, Imbalanced data, K-fold Cross Validation, Leave One Out*

I. PENDAHULUAN

Regresi logistik merupakan analisis regresi yang digunakan dalam pemodelan klasifikasi. Dalam analisisnya, regresi logistik akan menghasilkan sebuah model. Model yang dibentuk perlu dinilai akurasinya untuk mendapatkan model yang layak diterapkan. Salah satu cara menilai akurasi model adalah dengan melihat laju galat. Laju galat dilihat dari perbandingan banyaknya galat dengan keseluruhan data yang digunakan dalam analisis. Perhitungan laju galat ini dilakukan dengan mempertimbangkan bahwa suatu metode mungkin menjadi yang paling baik ketika digunakan dalam memprediksi suatu gugus data, tetapi metode tersebut belum tentu dapat memprediksi gugus data yang berbeda dengan baik.. Oleh karena itu, penting untuk memprediksi laju galat gugus data tersebut sehingga mendapatkan metode yang layak. Metode yang dapat digunakan dalam memprediksi laju galat salah satunya adalah *cross validation* (CV). Metode ini membagi data menjadi data latih (data *training*) dan data tes (data *testing*) (Molinaro dkk, 2005).

CV memiliki beberapa algoritma pembelajaran dan yang paling umum digunakan adalah *leave one out* (LOO), *hold out* dan *k-fold cross validation*. Ketiga algoritma tersebut memiliki perbedaan dalam pembagian data *training* dan data *testing*. LOO tidak membagi data secara acak karena pembagian datanya hanya meninggalkan satu pengamatan untuk data *testing* sehingga tidak perlu dilakukan pengacakan (Wong, 2015). Refaeilzadeh, dkk (2016) dalam artikelnya mengatakan bahwa LOO menghasilkan estimasi akurasi hampir tidak bias, tetapi memiliki variansi tinggi yang mengarah kepada penelitian yang tidak dapat diandalkan. LOO umumnya digunakan pada penelitian jumlah amatan kecil, karena ketika jumlah amatan besar maka akan membutuhkan cukup banyak waktu untuk melakukan analisis. *Hold out* merupakan algoritma prediksi laju galat yang paling sederhana. Pembagian data pada *hold out* mengakibatkan ada kemungkinan data yang penting pada data *training* masuk kedalam data *testing*, sehingga mempengaruhi kinerja algoritma. Menurut Refaeilzadeh, dkk (2016), *k-fold cross validation* adalah algoritma prediksi laju galat yang membagi data ke dalam beberapa kelompok. Dengan mengelompokkan data kedalam beberapa kelompok akan memudahkan dalam proses perhitungan. *K-fold* lebih baik dari segi komputasi dibanding dengan LOO.

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/86

Selain itu, variansi data yang dihasilkan pada *k-fold* cenderung lebih rendah. Kelemahan dari *k-fold* adalah data dengan ukuran sampel yang kecil tidak cocok digunakan.

Gugus data riil hasil pengumpulan data kebanyakan merupakan data tidak seimbang. Data tidak seimbang (*imbalanced*) adalah data yang memiliki jumlah kelas amatan yang berbeda. Data tidak seimbang memberi tantangan tersendiri dalam pengklasifikasian. Data tidak seimbang jika diklasifikasikan dengan benar akan memberikan nilai prediksi yang lebih akurat (Maalouf & Trafalis, 2011). Ketidakseimbangan data berdampak pada hasil prediksi yang tidak stabil. Hasil prediksi akan cenderung mengarah pada kelas mayoritas dan mengabaikan kelas minoritas. ketidakseimbangan data juga berdampak pada *error* yang dihasilkan. Semakin tidak seimbang proporsi kelas data maka *error* yang dihasilkan akan mengarah kepada proporsi kelas minoritas sehingga memberikan hasil *error* yang lebih kecil. Namun hal ini bukan mengartikan bahwa semakin tidak seimbang suatu data maka semakin baik (Chawla, 2010).Selain ketidakseimbangan data, korelasi juga memberikan dampak kepada hasil prediksi. Pada regresi logistik variabel bebas tidak boleh memiliki korelasi yang tinggi karena akan mengakibatkan model yang menjadi bias dan selang kepercayaan yang dihasilkan sangat lebar (Courvoisier, dkk, 2011). Rumusan masalah pada penelitian ini adalah bagaimana kinerja metode prediksi laju galat LOO*, hold out* dan *k-fold cross validation* terhadap data tidak seimbang dengan proporsi kelas data yang berbeda serta metode prediksi laju galat apa yang cocok diterapkan pada regresi logistik biner untuk data tidak seimbang. Tujuan dari penelitian ini adalah mengidentifikasi kinerja masing-masing algoritma prediksi laju galat pada data tidak seimbang dengan proporsi kelas data yang berbeda, membandingkan kinerja algoritma prediksi laju galatsehingga dapat mengetahui algoritma prediksi laju galat yang paling cocok diterapkan pada model regresi logistik biner dengan data tidak seimbang.

II. METODE PENELITIAN

Penelitian ini menggunakan data bangkitan yang diperoleh dari data hasil simulasi pada *Rstudio version 4.1.2.* Data yang dibangkitkan sebanyak 145 gugus data. Selanjutnya, untuk tahapan pembangkitan data dan tahapan analisis akan dijelaskan pada bagian berikut.

A. Membangkitkan Data

Data yang dibangkitkan akan diperhatikan perbandingan nilai median dan variasi dari *error rate* yang diperoleh. Dalam pemodelan regresi logistik biner variabel terikat berbentuk biner dengan nilai 0 atau 1. Sedangkan untuk variabel bebasnya dibangkitkan dengan beberapa pengaturan yang diulas yaitu, jumlah variabel bebas, perbedaan rataan populasi dari sampel berasal, dan korelasi antara variabel untuk kasus data bivariat dan multivariat.

Pengaturan jumlah variabel bebasterdiri atas satu variabel bebas (univariat), dua variabel bebas (bivariat) dan tiga variabel bebas (multivariat). Pada kasus data bivariat dan multivariat dikaji untuk memungkinkan simulasi yang melibatkan struktur korelasi dan struktur rataan yang berbeda. Untuk kasus univariat data dibangkitkan dari distribusi normal N(μ,σ) dengan perbedaan rataan populasi pada Tabel 1 berikut.

Tabel 1. Ketentuan nilai rataan populasi data univariat

Kasus	$\boldsymbol{\mu^{(1)}}$	$\boldsymbol{\mu^{(2)}}$
1	0	1
2	0	2

Pada Tabel 1, kedua kasus data dibangkitkan dengan distribusi normal dengan variansi satu (σ = 1). Sementara itu, untuk kasus data bivariat dan multivariat perbedaan rataan populasi diuraikan pada Tabel 2 berikut.

Tabel 2. Ketentuan nilai rataan populasi data bivariat dan multivariat

Jumlah Variabel	Kasus	Struktur Rataan Populasi $\underline{\mu}^{(1)}$	$\underline{\mu}^{(2)}$
Bivariat	1	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}1\\1\end{pmatrix}$
	2	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}1\\0\end{pmatrix}$
	3	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}2\\2\end{pmatrix}$
	4	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}2\\0\end{pmatrix}$

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/86

ISSN(*Online*): 2985-475X

Jumlah Variabel	Kasus	Struktur Rataan Populasi $\underline{\mu}^{(1)}$	$\underline{\mu}^{(2)}$
Multivariat	1	$\begin{pmatrix}0\\0\\0\end{pmatrix}$	$\begin{pmatrix}1\\1\\1\end{pmatrix}$
	2	$\begin{pmatrix}0\\0\\0\end{pmatrix}$	$\begin{pmatrix}1\\1\\0\end{pmatrix}$
	3	$\begin{pmatrix}0\\0\\0\end{pmatrix}$	$\begin{pmatrix}1\\0\\1\end{pmatrix}$

Pada Tabel 2 untuk kasus 1 pada data bivariat kedua variabel penjelas dinamakan variabel relevan karena variabel memuat informasi kelompok populasi yang berbeda begitu juga dengan kasus 3. Sedangkan untuk kasus 2 dan kasus 4 variabel pertama merupakan variabel relevan dan variabel kedua merupakan variabel irrelevan. Dikatakan variabel irrelevan karena kedua kelompok populasi berasal memiliki informasi yang sama. Sementara itu, pada data multivariat untuk kasus 1 ketiga variabel bebas merupakan variabel relevan. Sedangkan pada kasus 2 variabel pertama dan kedua merupakan variabel relevan dan variabel ketiga merupakan variabel irrelevan. Begitu juga dengan kasus 3 yang merupakan variabel relevan adalah variabel pertama dan ketiga, sedangkan variabel kedua merupakan variabel irrelevan. Selanjutnya untuk struktur korelasi pada data bivariat dapat dilihat pada Tabel 3 berikut.

Tabel 3. Ketentuan korelasi pada data bivariat

Korelasi	Struktur Korelasi
A	$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$
B	$\begin{bmatrix}1 & 0.5\\0.5 & 1\end{bmatrix}$
C	$\begin{bmatrix}1 & 0.9\\0.9 & 1\end{bmatrix}$

Pada Tabel 3 korelasi A menunjukkan kasus data dengan tanpa korelasi, korelasi B menunjukkan kasus data dengan korelasi sedang dan korelasi C menunjukkan kasus data dengan korelasi tinggi. Kemudian untuk pembangkitan data pengaturan struktur korelasi digabungkan dengan pengaturan struktur rataan populasi sehingga dapat ditelaah kondisi data dengan variabel relevan yang tidak memiliki korelasi, memiliki korelasi sedang atau korelasi tinggi dan kondisi data terdapat variabel relevan dan variabel irrelevan yang tidak memiliki korelasi, memiliki korelasi sedang atau korelasi tinggi. Begitu juga dengan data multivariat juga akan dilakukan hal yang sama. Untuk data multivariat, korelasi yang digunakan dapat dilihat pada Tabel 4 berikut.

Tabel 4. Ketentuan korelasi pada data multivariat

Korelasi	Struktur Korelasi
A	$\begin{bmatrix}1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\end{bmatrix}$
B	$\begin{bmatrix}1 & 0.5 & 0\\0.5 & 1 & 0\\0 & 0 & 1\end{bmatrix}$
C	$\begin{bmatrix}1 & 0.9 & 0\\0.9 & 1 & 0\\0 & 0 & 1\end{bmatrix}$
D	$\begin{bmatrix}1 & 0.5 & 0.5\\0.5 & 1 & 0.5\\0.5 & 0.5 & 1\end{bmatrix}$
E	$\begin{bmatrix}1 & 0.9 & 0.9\\0.9 & 1 & 0.9\\0.9 & 0.9 & 1\end{bmatrix}$

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/86

ISSN(*Online*): 2985-475X

Penelitian ini menggunakana data tidak seimbang. Data tidak seimbang adalah data yang memiliki jumlah proporsi kelas amatan yang berbeda. Ketidakseimbangan data juga akan berpengaruh terhadap hasil *error rate*. Untuk pengaturan proporsi ketidakseimbangan data digunakan pengaturan rasio amatan yang diuraikan pada Tabel 5 berikut. Pada masing-masing perbandingan amatan akan diterapkan pengaturan struktur rataan populasi dan struktur korelasi untuk data bivariat dan multivariat.

Tabel 5. Rasio jumlah amatan

Pengaturan	Rasio
1	50:50
2	60:40
3	70:30
4	80:20
5	90:10

B. Membangun Model Regresi Logistik Biner

Setelah melakukan pembangkitan data, selanjutnya adalah membangun model regresi logistik biner. Regresi logistik merupakan analisis regresi yang variabel terikatnya bersifat kategorik. Regresi logistik dikatakan biner karena variabel Y terdiri atas dua kategori (biner), diamana kategori berupa sukses dan gagal. Kategori sukses dilambangkan dengan angka "1" dan kategori gagal dilambangkan dengan angka "0"(Hosmer & S.Lemeshow, 2013).

Abonazel dan Ibrahim (2018) mendefinisikan bahwa regresi logistik merupakan kasus khusus dari model linier umum sehingga dikatakan mirip dengan regresi linier. Perbedaannya adalah distribusi yang digunakan dalam regresi logistik adalah distribusi Bernouli sedangkan pada regresi linier distribusi yang digunakan adalah distribusi normal dan pada regresi logistik nilai prediksi berbentuk probabilitas sedangkan pada regresi linier nilai prediksi berbentuk kontinu. Model dari regresi logistik dengan k variabel adalah sebagai berikut.

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)} \quad (1)$$

Nilai probabilitas $\pi(x)$ berkisar antara nol dan satu. Sedangkan nilai linier π terhadap nilai x berkisar antara $(-\infty, +\infty)$ sehingga untuk memungkinkan apapun nilai x yang dimiliki berkisar antara nol dan satu perlu dilakukan transformasi pada model regresi logistik dengan transformasi logit. Hasil dari transformasi logit adalah sebagai berikut.

$$g(x) = ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \quad (2)$$

Dimana:

$\pi(x)$: probabilitas sukses dengan nilai probabilitas $0 \leq \pi(x) \leq 1$

k : banyak variabel bebas

Nilai probabilitas $\pi(x)$ merupakan nilai prediksi pada regresi logistik, namun pada regresi logistik biner nilai prediksinya adalah g(x). Nilai prediksi disimbolkan dengan $\hat{y}$ dan nilai asli disimbolkan dengan y. Adapun syarat dari regresi logistik menurut Abonezel dan Ibrahim (2011) adalah variabel terikat dan variabel bebas tidak membutuhkan hubungan linier sehingga asumsi multikolinieritas tidak ada. Tidak dibutuhkan asumsi *error varians* (residual). Kehomogenan data tidak diperlukan sehingga asumsi homoskedastisitas tidak diperlukan. variabel terikat bersifat dikotomi atau memiliki dua kategori. Variabel bebas tidak perlu diubah ke dalam bentuk skala rasio atau interval dan variabel bebas tidak harus memiliki keragaman yang sama antara kelompok variabel, sehingga variabel bebas bersifat saling bebas atau eksklusif.

Menurut Courvoisier, dkk, (2011), pada regresi logistik variabel bebas tidak boleh memiliki korelasi yang tinggi karena akan mengakibatkan multikolinieritas. Korelasi yang tinggi akan menyulitkan dalam estimasi koefisien model sehingga menghasilkan estimasi koefisien yang tidak stabil dan tidak dapat diandalkan serta sulit untuk menginterpretasikannya. Hal ini juga mengakibatkan model yang dihasilkan biasnya lebih tinggi. Selain itu, korelasi yang tinggi juga menyebabkan selang kepercayaan yang dihasilkan sangat lebar sehingga sangat sulit untuk menolak hipotesis nol.

C. Melakukan Perbandingan Prediksi Laju Galat

Setalah model dibentuk, langkah selanjutnya adalah melakukan prediksi laju galat dengan membandingkan kinerja algoritma prediksi lau galat LOO, *k-fold*, dan *hold out*. Galat atau *error* merupakan selisih antara nilai yang kita harapkan dengan nilai sebenarnya (Purwati & Erawati, 2020). Pada klasifikasi untuk mengukur galat atau *error* menggunakan fungsi indikator dengan rumus sebagai berikut.

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/86

ISSN(*Online*): 2985-475X

$$Err_i = I(y_i \neq \hat{y}_i)$$

Dimana,

$$I = \begin{cases} 1 & y_i \neq \hat{y}_i \\ 0 & y_i = \hat{y}_i \end{cases}$$

Salah satu metode yang dapat digunakan untuk memprediksi laju galat adalah metode CV. Menurut James, dkk, (2013), CV merupakan teknik penguji keefektifan dari model yang dibentuk dengan melakukan penyusunan ulang (*resampling*) pada data untuk membaginya menjadi dua bagian yaitu data *training* dan data *testing*. Data *training* akan dipakai untuk melatih model sehingga model dapat memahami pola pada data sedangkan untuk melakukan validasi terhadap model tersebut, akan digunakan data *testing* sebagai pengujiannya. CV melakukan pembagian data berulang kali menjadi dua bagian, dimana bagian pertama digunakan untuk melatih model dan bagian kedua digunakan untuk meguji model. CV mengasumsikan bahwa data *training* dan data *testing* bersifat *independen* (saling bebas). Tujuan dari CV adalah melakukan pengujian terhadap suatu model yang telah dibentuk dari data yang tersedia dengan algoritma yang terdapat pada CV kemudian membandingkan kinerja masing-masing algoritma sehingga mendapatkan algoritma yang cocok digunakan pada analisis data (Refaeilzadeh, dkk, 2016). CV memiliki beberapa algoritma pembelajaran dan yang paling umum digunakan adalah *leave one out, hold out* dan *k-fold cross validation*. Ketiga algoritma tersebut akan dibandingkan untuk mendapatkan algoritma yang paling cocok diterapkan pada pemodelan regresi logistik biner untuk data tidak seimbang.

Leave one out (LOO) merupakan salah satu metode prediksi laju galat yang membagi data menjadi dua bagian yaitu data *training* dan data *testing* berdasarkan jumlah pengamatan, dimana untuk data *training* berisi sebanyak N-1 pengamatan dan satu pengamatan menjadi data *testing.* Proses perhitungan dapat diulang sebanyak n kali dimana selalu meninggalkan satu pengamatan untuk data *testing* (James, dkk, 2013). Namun pembagian ini memiliki kelemahan dalam segi komputasi, data yang digunakan berjumlah sangat besar maka akan menimbulkan permasalahan dalam komputasi. Tingkat akurasi yang diperoleh LOO hampir tidak bias tetapi variansi yang dihasilkan tinggi (Refaeilzadeh, dkk, 2016). Untuk menghitung nilai *error rate* pada algoritma LOO dapat menggunakan rumus sebagai berikut.

$$\hat{E}^{LOO} = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i) \tag{3}$$

Hold out merupakan salah satu algoritma yang ada pada CV. *Hold out* membagi dua data secara acak dimana dua pertiga data dijadikan sebagai data *training* dan sepertiga lainnya digunakan untuk data *testing*. Pembagian data pada *hold out* dilakukan pada masing-masing kelas. *Hold out* merupakan algoritma yang lebih sederhana dibandingkan dengan algoritma lainnya (James, dkk, 2013). Namun, *hold out* memiliki kelemahan dalam pembagian data. Tidak semua data berkesempatan menjadi data *testing* sehingga berkemungkinan tidak semua data yang menjadi data *testing* dapat memprediksi dengan baik ataupun yang menjadi data *training* dapat memprediksi dengan baik sehingga hasil yang diperoleh sangat bergantung pada pemisahan data. Hal ini sebenarnya dapat diatasi dengan melakukan perulangan dalam pengacakan data dan pembagian data namun karena pembagiannya secara acak masih ada kemungkinan data tidak pernah menjadi data *testing* (Refaeilzadeh, dkk, 2016). Untuk menghitung nilai *error rate* pada algoritma *hold out* dapat menggunakan rumus sebagai berikut.

$$\hat{E}^{HO} = \frac{1}{n_{test}}\sum_{i=1}^{n_{test}} I(y_i \neq \hat{y}_i) \tag{4}$$

K-fold cross validation merupakan salah satu metode prediksi laju galat yang membagi data ke dalam k kelompok dengan ukuran sampel yang sama. Untuk data *training* menggunakan sebanyak k-1 kelompok sedangkan sisa satu kelompok digunakan untuk data *testing. K-fold* membagi data secara acak dan mengelompokkan data tersebut kedalam beberapa kelompok dengan jumlah pengamatan masing-masing kelompok sama. Banyak kelompok yang biasa digunakan adalah 5 atau 10 kelompok (*fold*) (James, dkk, 2013). Untuk menghitung nilai *error rate* pada algoritma *k-fold* dengan 5 kelompok dapat menggunakan rumus sebagai berikut.

$$\hat{E}^{CV} = \frac{1}{5}\sum_{k=1}^{5}\frac{1}{n_k}\sum_{i=1}^{n_k} I(y_i \neq \hat{y}_i) \tag{5}$$

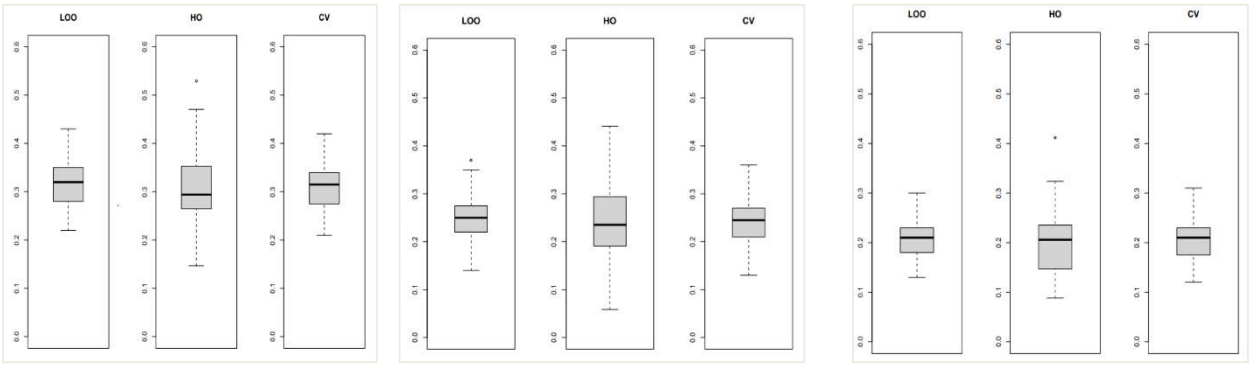
III. HASIL DAN PEMBAHASAN

Penelitian ini ditujukan untuk melihat algoritma prediksi laju galat terbaik yang cocok digunakan untuk metode regresi logistik biner data tidak seimbang. Kinerja algoritma prediksi laju galat dapat dilihat dari variasi *error rate*. Variasi *error rate* dapat dilihat berdasarkan boxplot yang dihasilkan. Semakin kecil boxplot yang dihasilkan maka

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/86

variasi *error rate* juga semakin kecil. Selain itu juga dapat melihat pengaruh antara perbedaan rataan populasi, korelasi dan ketidakseimbangan data terhadap prediksi laju galat.

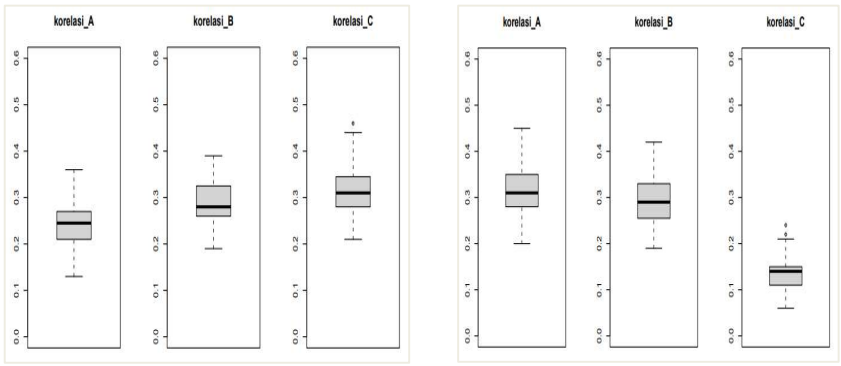
Pada Gambar 1 dapat dilihat bahwa perbedaan hasil *error rate* pada setiap jenis data. Nilai *error rate* semakin kecil ketika jumlah variabel bebas semakin banyak. Hal ini berarti bahwa data multivariat memiliki nilai *error rate* yang kecil. Pada Gambar 1 dapat dilihat algoritma *hold out* merupakan algoritma yang memiliki variasi *error rate* yang paling besar diantara ketiga algoritma prediksi laju galat. Namun nilai median *error rate* yang dihasilkan algoritma *hold out* cenderung lebih kecil dibandingkan dengan algortima lain. Hal ini dikarenakan algoritma *hold out* cenderung tidak stabil dalam memprediksi laju galat karena jumlah data *training* pada *hold out* lebih sedikit dibandingkan algoritma lain. Algoritma LOO dan *k-fold* memiliki kinerja yang hampir sama dalam memprediksi laju galat. Variasi *error rate* yang dihasilkan hampir sama besar namun jika diperhatikan lebih jelas lagi variasi *error rate* algoritma *k-fold* lebih kecil dibandingkan algoritma LOO.



(a) (b) (c)

Gambar 1. Hasil *error rate*algoritma prediksi laju galat pada data(a) univariat (b) bivariat (c) multivariat dengan kasus 1

Pada data univariat dan bivariat, algoritma *k-fold* lebih unggul dibandingkan algoritma LOO. Hal ini dikarenakan data yang digunakan tidak terlalu besar, sehingga pada algoritma LOO variasi *error rate* yang dihasilkan cenderung lebih kecil. Tetapi pada data multivariat ada beberapa kasus data yang algoritma LOO lebih unggul daripada algoritma *k-fold*, namun secara keseluruhan algoritma *k-fold* lebih baik dalam melakukan prediksi laju galat dibandingkan algoritma LOO. Sehingga dapat dikatakan bahwa algoritma yang cocok digunakan untuk memprediksi laju galat pada pemodelan regresi logistik biner dengan data tidak seimbang adalah algoritma *k-fold cross validation*. Untuk pengaturan berikutnya mendapatkan hasil yang sama bahwa algoritma prediksi laju galat yang cocok adalah algoritma *k-fold cross validation*.

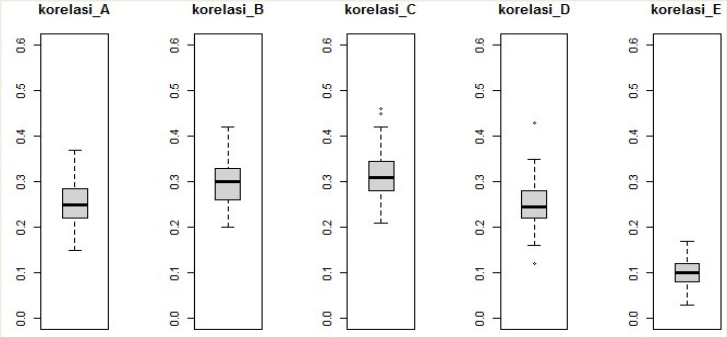


(a) (b)

Gambar 2. Hasil *error rate* algoritma *k-fold* dengan yang berkorelasi (a) sesama variabel relevan (b) variabel relevan dan variabel irrelevan

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/86

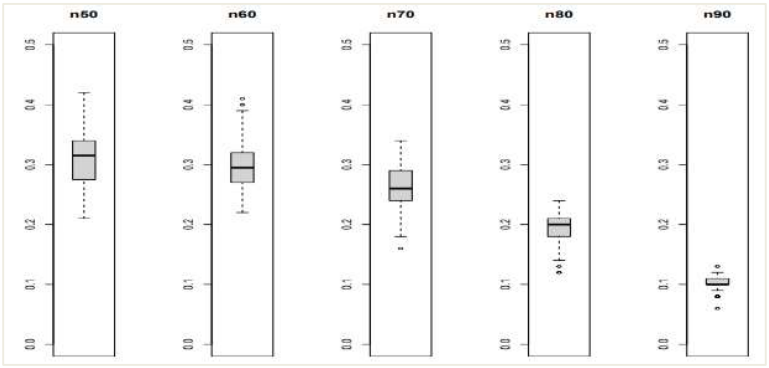
Pada data bivariat dan multivariat dapat dilihat perbedaan nilai prediksi laju galat pada kasus korelasi yang berbeda. Data yang memuat variabel relevan pada semua variabelnya ketika berkorelasi akan meningkatkan hasil *error rate* jika nilai korelasinya semakin besar, seperti yang dapat dilihat pada Gambar 2(a). Pada Gambar 2(b) variabel pertama merupakan variabel relevan dan variabel kedua merupakan variabel irrelevan. Ketika kedua variabel tersebut berkorelasi hasil *error rate* semakin turun ketika korelasi semakin tinggi. Hasil yang sama juga didapatkan pada pengaturan yang lainnya.



Gambar 3. Hasil *error rate* algoritma *k-fold* dengan korelasi antara variabel relevan dengan variabel irrelevan pada data multivariat kasus 2

Pada data multivariat, variabel bebas yang berkorelasi adalah dua variabel yaitu variabel pertama dengan variabel kedua dan tiga variabel. Gambar 3 merupakan hasil *error rate* algoritma *k-fold* yang berkorelasi, dimana variabel pertama dan kedua merupakan variabel relevan dan variabel ketiga merupakan variabel irrelevan. Pada Gambar 3 dapat dilihat bahwa ketika yang berkorelasi dua variabel dengan kedua variabel tersebut merupakan variabel relevan maka *error rate* yang dihasilkan akan meningkat. Namun ketika ada satu variabel irrelevan pada data tersebut maka *error rate* yang dihasilkan akan menurun. Pada data yang berkorelasi tiga variabel juga sama. Ketika semua variabel yang berkorelasi merupakan variabel relevan maka akan terjadi peningkatan terhdap nilai *error rate*, namun jika terdapat salah satu variabel merupakan variabel irrelevan maka *error rate* yang dihasilkan akan menurun.

Data yang dibangkitkan dengan proporsi kelas data yang berbeda menghasilkan *error rate* yang berbeda seperti dapat dilihat pada Gambar 4.



Gambar 4. Perbandingan hasil *error rate* dengan jumlah kelas amatan yang berbeda pada algoritma *k-fold* data univariat

Pada Gambar 4 dapat dilihat bahwa data dengan proporsi seimbang akan menghasilkan variasi *error rate* yang lebih besar dan nilai *error rate* yang dihasilkan juga besar. Namun ketika data dengan proporsi yang sangat tidak seimbang cenderung akan menghasilkan variasi *error rate* yang kecil dan nilai *error rate* yang dihasilkan juga kecil. Namun hal ini bukan berarti semakin tidak seimbang suatu data maka dikatakan baik tetapi ini merupakan sebuah kesalahan. Pada data tidak seimbang, kesalahan prediksi atau *error* sering terjadi pada kelas minoritas sehingga *error rate* yang dihasilkan cenderung lebih kecil. Seperti yang terdapat pada Gambar 4 ketika perbandingan proporsi kelas

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/86

data seimbang dengan n = 50:50, hasil *error rate* menyebar di antara selang 0.2 sampai 0.45, tetapi pada data dengan proporsi kelas yang sangat tidak seimbang dengan perbandingan kelas n = 80:20 hasil *error rate* berada hanya disekitar angka 0.2 begitu juga dengan perbandingan kelas n = 90:10 hasil *error rate* hanya berada disekitar angka 0.1.

IV. KESIMPULAN

Algoritma LOO dan *k-fold* menunjukkan kinerja yang hampir sama dalam memprediksi laju galat, tetapi algoritma yang memiliki variasi *error rate* yang terkecil adalah algoritma *k-fold*. Oleh karena itu, algoritma *k-fold* cocok digunakan untuk memprediksi laju galat dalam memodelkan regresi logistik biner dengan data tidak seimbang. Keberadaan korelasi antar variabel pada gugus data *bivariat* dan *multivariat* mempengaruhi hasil prediksi laju galat. Korelasi antara sesama variabel *relevan* menunjukkan hasil yang berbeda dengan korelasi antara variabel *relevan* dan variabel *irrelevan*. Ketika berkorelasi sesama variabel *relevan*, hasil prediksi laju galat semakin besar jika nilai korelasi semakin besar, sedangkan ketika berkorelasi antara variabel *relevan* dan *irrelevan* hasil prediksi laju galat menunjukkan penurunan ketika nilai korelasi semakin besar. Ketidakseimbangan data memberi pengaruh terhadap *error rate* yang dihasilkan oleh ketiga algoritma. Semakin tidak seimbang data, maka *error rate* yang dihasilkan semakin kecil. Hal ini dikarenakan semakin tidak seimbang suatu data, maka hasil *error rate* akan mendekati proporsi kelas minoritas. Sehingga *error rate* yang dihasilkan cenderung lebih kecil dan tidak adanya penanganan yang dilakukan untuk mengatasi ketidakseimbangan data. Diharapkan untuk penelitian selanjutnya dapat mengembangkan penelitian baru dengan menambahkan metode penanganan ketidakseimbangan data agar mendapatkan hasil yang lebih baik.

DAFTAR PUSTAKA

Abonazel, M., & Ibrahim, M. (2018). On Estimation Methods for Binary Logistic Regression Model with Missing Values. *International Journal of Mathematics and Computational Science, Vol.4 No.3*, 79-85. Retrieved from http://www.aiscience.org/journal/ijmcs

Chawla, N. (2010). Data Mining for Imbalanced Datasets: An Overview. *Data Mining and Knowledge Discovery Handbook*, 875-886.

Courvoisier, D., Combescure, C., Agoritsas, T., Gayet-Ageron, A., & Pemeger, T. (2011). Performance of Logistic Regression Modeling: Beyond The Number of Events per Variable, The Role of Data Structure . *Journal of Clinical Epidemiology*, 993-1000.

Hosmer, D., & S.Lemeshow. (2013). *Applied Logistic Regression.Edisi ke-3 John Wiley and Sons Inc.* Canada: New Jersey.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Application in R.* New York: Springer.

Maalouf, M., & Trafalis, T. (2011). Robust weighted kernel logistic regression in imbalanced and rare events data. *Computational Statistics and Data Analysis*, 168-183.

Molinaro, A., Richard, S., & Pfeiffer, R. (2005). Prediction Error Estimation: a Comparison of Resampling Methods. *BIOINFORMATICS, Vol 21 no 15*, 3301-3307. doi:10.1093/bioinformatics/bti499

Purwati, N., & Erawati, N. (2020). *Pengantar Metode Numerik.* Lumajang: Klik Media.

Refaeilzadeh, P., Tang, L., & Liu, H. (2016). Cross Validation. *Encyclopedia of Database Systems*, DOI:10.1007/978-1-4899-799-3_565-2.

Wong, T.-T. (2015). Performance Evaluation of Classification Algorithms by K-Fold and Lesve-One-Out Cross Validation. *attern Recorgnition*, 1-8.

DOI: https://doi.org/10.24036/ujsds/vol1-iss4/86